DOCUMENT RESUME

ED 248 695 FL 014 535

AUTHOR Crawford, Gary D.; And Others
TITLE Oral Language Proficiency Testing at the Foreign Service Institute. An Update--1983.
INSTITUTION Foreign Service (Dept. of State), Washington, D.C. Foreign Service Inst.
PUB DATE 83
NOTE 12p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Standards; *Adult Students; Difficulty Level; Evaluation Criteria; Interviews; Language Proficiency; *Second Language Instruction; *Test Construction; *Test Use
IDENTIFIERS *Foreign Service Institute DC; *Oral Proficiency Testing

ABSTRACT

The Foreign Service Institute (FSI) has been engaged in oral language proficiency testing theory and practice for more than 20 years. The FSI test has been consistent during this time in format, evaluation criteria, performance standards, and level definitions. Current concerns about the degree of standardization of the format and the strength of the inferences drawn from the test motivated the efforts to improve it. It was decided to revise: (1) the evaluation criteria by adding "discourse competence," reconstituting "structural precision," expanding "lexicalization," redefining "fluency," and de-limiting "comprehension;" (2) the performance standards, resulting in generic standards to describe performance with regard to six factors (blocking, dysfunctional, intrusive, acceptable, successful, and superior); and (3) the test format, resulting in standard stimuli common to tests in all languages and now consisting of three separate parts (a conversation, an interview task, and a briefing). (Author/MSE)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED248695

# ORAL LANGUAGE PROFICIENCY TESTING AT THE FOREIGN SERVICE INSTITUTE

## An Update -- 1983

Gary D. Crawford, H. David Argoff, Marianne L. Adams*


## ABSTRACT

The Foreign Service Institute has been engaged in oral language proficiency testing theory and practice for more than 20 years. The "FSI TEST" has been consistent during this time span in format, evaluation criteria, performance standards and level definitions.

Current concerns about the degree of standardization of the format and the strength of the inferences which are drawn from the test motivated the efforts to improve it.

It was decided to revise : (1) the evaluation criteria,, (2) the performance standards and (3) the test format. The resulting changes are these: revised five evaluation factors--comprehension, discourse competence, fluency, lexicalization, and structural precision; six generic standards which describe performance with regard to each of the five factors: blocking, dysfunctional, intrusive, acceptable, successful and superior. The format, with standard stimuli common to tests in all languages now consists of three separate parts: a conversation, an interview task, and a briefing.


---

* H. David Argoff and Marianne L. Adams are FSI's test development committee working under the direction of Gary Crawford.

FL 014 535

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Gary D. Crawford

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

# ORAL LANGUAGE PROFICIENCY TESTING AT THE FOREIGN SERVICE INSTITUTE

## An Update

For many years, the Foreign Service Institute's Language School has played a major role in developing the theory and practice of oral language proficiency testing. The Oral Interview Proficiency Test and the S-Scale definitions continue to be spoken of as the "FSI testing system". Although much of the early work was done by FSI personnel, the "FSI test" has gone well beyond the walls of our Institute. With various refinements, the system has been adopted by all U.S. Government testing services, by many educators and testers throughout the U.S., and by military training institutes of the NATO countries.

The system has remained remarkably intact over the years. It consists of four main components:

1. TEST FORMAT: A face-to-face 30-60 minute encounter between the examinee and a native speaker (the "interviewer"), observed by a testing specialist (the "examiner") certified to award scores. Test events are not rigidly controlled, consisting largely of conversation at the direction of the interviewer.

2. EVALUATION CRITERIA: The features of the examinee's performance that are considered indicative of proficiency, and from which a score is derived. These features have been grouped under five "factors": Accent, Grammar, Fluency, Vocabulary, and Comprehension.

3. PERFORMANCE STANDARDS: A scale of excellence for each factor, used to describe a performance in order to arrive at a global S-score.

4. LEVEL DEFINITIONS: A scale of overall ("global") proficiency in functional terms, from S-0 (no proficiency) to S-5 (proficiency equivalent to that of an educated native speaker).

The strength of this system is its flexibility. The nature of the interview can be adjusted to fit the circumstances, and experienced testers agree on global S-scores despite disagreements on the significance of specific points. For many years this system has been the best option for assessing communicative competence, when that--rather than course achievement--is what is wanted.

Changes have been made and continue to be made in the system. Various institutions have adopted and adapted the system in order to serve their particular evaluation requirements. In November, 1983, the Interagency Language Roundtable, comprising representatives of government agencies and bureaus which have an interest in foreign language use and/or training, completed a three-year project to rewrite the language proficiency descriptions of the basic levels and, for the first time, to give explicit definitions for the intermediate points in the 6-point Scales. Recently, the FSI Language School implemented several important revisions of its own.

This paper is a report on the revised FSI testing system.

## The Concerns

All tests are inferential. The only truly direct assessment is on-the-job observation which, in our case, is out of the question. By definition, a "test" is an attempt to make valid inferences from a short-term demonstration. The validity of the score depends on the strength of the inferences:

> "I observed the examinee do A, B, and C well, but have difficulty with D and E. I infer from these observations that in real life he can handle X-type situations fairly successfully, but not Y-type situations."

FSI was concerned that the test was too inferential. That is, judgments were sometimes based on too little evidence or too shaky evidence. There was a tendency to make "inferential short-cuts": "People who make errors in noun plurals can't be above an S-2".

Conversely, there was sometimes evidence of communicative competence that could not be described solely in terms of the five traditional factors. This was particularly apparent in special cases such as "rusty" native speakers, whose ability to communicate was impressive despite error-laden speech.

CONCERN #1: STRENGTHEN THE INFERENCES THAT LEAD TO S-SCORES.

Our second concern was to ensure that the same score in different languages means the same thing. Can an S-3 in Chinese do what an S-3 in Spanish can? More? Less? How can we be sure? The S-Scale is defined in functional terms--what people can do at each level--but these functional scores are derived from a non-functional assessment. Testers tend to notice "correctness" (errors or lack of errors) rather than functional "effectiveness".

To ascertain effectiveness it is necessary to pose communicative tasks of some sort and observe the examinee's attempts to perform them. Moreover, to ensure comparability between languages, all examinees should confront the same tasks.

CONCERN #2: DEVELOP A STANDARDIZED TASK-BASED TEST FORMAT

## The Revisions

For our assessments to include functional effectiveness as well as linguistic correctness, we had to revise the evaluation criteria and the performance standards associated with them. The test format had to be changed for two reasons: (1) to ensure that examinees had the opportunity to demonstrate their functional capabilities, and (2) to ensure that examinees in different languages had the same opportunities.

The level definitions, already expressed in terms of functional effectiveness, were left unchanged. Indeed, all revisions were designed to be consistent with the well-established S-Scale and the thousands of S-Scores awarded over the years. Our goal was to improve the validity of our test scores, not to change test itself.

The three revisions are described below.

1. EVALUATION CRITERIA. Testers listen to everything, but what should they listen for? We wanted our testers to include in their judgements everything that contributes to or detracts from the effectiveness of communication.

The is no dearth of possible evaluation factors. On the contrary, there are too many, and they tend to overlap or subsume each other. Since there is a finite limit to the number of factors that can be noticed simultaneously, we resolved to stay with five.

The following outline indicates how and why the traditional five factors were altered.

o We wanted to observe features of the examinee's performance such as "organization", "turn-taking", "conversational repair", "linking/ cohesion", and other characteristics of skillful discourse.

NEW FACTOR: Discourse Competence

o We wanted to put non-discretionary rule-governed matters of correctness (i.e., accuracy rather than appropriateness) together under one factor. Dropping "accent" and "grammar", we established a composite factor that includes both morphology and phonology, as well as some aspects of word-ordering.

RE-CONSTITUTED FACTOR: Structural Precision

o We wanted to assess aptness of expression and appropriateness of usage, as well as breadth of vocabulary. We expanded "vocabulary" to include these concepts.

EXPANDED FACTOR: Lexicalization

o We wanted to observe how readily language comes to the examinee's mind, but not to count off for native-like pauses. We refined the definition of "fluency" along these lines.

RE-DEFINED FACTOR: Fluency

o We wanted to make it clear that the "comprehension" factor refers only to comprehending speech in face-to-face situations--where there is interaction between the speaker and hearer. Non-interactive comprehension, such as listening to speeches or the broadcast media, is not measured by this test.

DE-LIMITED FACTOR: Comprehension

SUMMARY For the purposes of proficiency evaluation, communicative competence consists (primarily) of five inter-related factors:

DISCOURSE COMPETENCE
STRUCTURAL PRECISION
LEXICALIZATION
FLUENCY
COMPREHENSION

2. PERFORMANCE STANDARDS. The revised factors serve to identify aspects of an examinee's performance that contribute to (or detract from) communicative effectiveness. To score a test, however, it is not enough for the examiner to say, "I now know the examinee's strengths and weaknesses." The examiner must also determine the extent of the examinee's abilities and disabilities.

To record the degree of effectiveness, examiners must refer to some set of performance standards. The nature of this scale is critical. Direct reference to the S-scale level definitions begs the question; seat-of-the-pants judgments are not stronger inferences. On the other hand, if the factors were scaled in abstract terms (say, along a semantic differential), or expressed in terms of an error count, then our revision would amount to little more than a change in terminology.

It was essential to develop a scale of effectiveness. Our revised performance standards set forth for each factor degrees of "impact" (pro or con) upon the success of the communication during the test.

There is no lessened concern at FSI for good grammar or good pronunciation. Any pattern of error that interferes with the success of communication lowers the score, as always. The revised standards, however, do allow examiners to lower scores for reasons other than incorrectness, and to raise scores for reasons other than correctness.

Specifically, the revised performance standards are as follows:

BLOCKING-- On a given factor, an examinee is rated "blocking" if the imperfections in his performance are such that they prevent the interaction from proceeding.

DYSFUNCTIONAL-- This rating is given when the imperfections force a temporary interruption in the interaction until significant "repairing" is accomplished.

INTRUSIVE-- The imperfections are clearly noticeable and disconcerting, but not otherwise problematic.

ACCEPTABLE-- The examinee's performance is such, on this factor, that the examiner observes no significant impediment or enhancement of effectiveness.

SUCCESSFUL-- The examinee's demonstrated facility on this factor is a significant contribution to the success of the communication.

SUPERIOR-- Control of this factor stands out as a major attribute of the examinee's success as a communicator.

These general definitions constitute a six-point scale of increasingly positive impact. For each factor, a very specific performance standard statement has been written for each degree of successfulness. In all the result is a 6 by 5 grid of thirty performance standard statements.

These statements relieve examiners of trying to describe the examinee's performance in their own words. Instead, they listen for evidence which will allow them to select the statement which best describes the degree of effectiveness they heard during the test.

The scoring grid contains only abbreviated statements. The full texts are contained elsewhere in our testing literature and form the basis of examiner-training. During a test, checks are placed in the approriate boxes on the grid. The scorers then refer to a scoring table to derive a raw score. A conversion chart converts the raw scores to S-scores.

Note that the scale of statements for a factor is not "smooth", with equal increments between the six points. The "distance" between intrusive and acceptable lexicalization, for example, is not necessarily the same as the distance between acceptable and successful lexicalization. The statements are, therefore, already "weighted" in consonance with existing level definitions.

3. FORMAT. The traditional format of the test--the "oral interview"--had to be better structured to ensure comparability between languages. Moreover, assessing communicative effectiveness requires richer samples of what examinees can and cannot do. Chit-chatting and question-answering alone show little about how well the examinee can manage a conversation or organize an oral presentation.

The ability of the examinee do these and other tasks common to the work of the foreign services is just what we want the test to reveal. Clearly then, the test should involve several very representative tasks which will permit testers to observe what we want them to assess.

Accordingly, a standard test format was developed, consisting of three representative tasks. They are:

1. a conversation
2. an interview
3. a briefing

The conversation is a warm-up period where tester and examinee exchange social/personal information. This gives an early indication of fluency, appropriateness of usage, among other impressions.

The interview has the examinee interviewing the tester to obtain his views on a subject of mutual interest. This task reveals how well the examinee can ask questions, make his agenda known, and understand what he is being told. The examinee provides the examiner with a point-by-point translation of the tester's views.

The briefing requires the examinee to deliver a 5-minute oral presentation on a professional topic and then respond to questions and comments from the tester. This task allows the examinee to demonstrate his ability to introduce and comment on new information, perform "linking", "transitions", "backtracking", "vocal/verbal underlining", "re-stating", and many other critical skills.

All tests in all languages follow this format. Test security is maintained by providing testers with numerous "kits" containing sets of topics from which the examinee may choose. These testing kits are periodically retired and re-combined.

## CONCLUSION

All FSI testers have now been re-trained to administer and score tests according to the new format. The change has been well received by both testers and students. The general reaction is that the revised format provides for a much better display of communication skills. The standard format is also fairly easy to learn. Gone is the necessity for surgical probing to determine whether or not the examinee has mastered the a particular grammatical feature. The tester interacts with the examinee in a reasonably natural way. All parties play themselves--there is no role-playing.

One spin-off benefit is that many more instructors can share the testing load at the Institute.

The re-training of examiners was a more ambitious and challenging effort, involving considerable tape-listening and discussion with our testing specialists. Examiner re-training continued throughout 1983. Our examiners who have completed the process have responded enthusiastically. They report that the new factors legitimize some of their observations which, in the past, they could not articulate and which could not influence the scoring.

We expect the average scores to remain very much the same--at least until our improved curricula take hold. Many individual scores, however, will come out higher or lower by the new scoring system. An extensive statisical analysis is planned to measure the effect of the revisions.

The following charts summarized the revised testing format and evaluation system and the retraining process being followed at the Foreign Service Institute:

6/13/83-updated 1-26-84

The FSI Oral Proficiency Test
The Revised System in a Nutshell

1. At the heart of a testing system are the evaluation factors -- those aspects of an examinee's performance which are significant and which contribute to the scores awarded. The set of evaluation factors constitutes the model of communicative competence on which the test is based. Below is a comparison of the five traditional evaluation factors with the five revised factors.

| TRADITIONAL FACTORS | REVISED FACTORS |
|---|---|
| • COMPREHENSION<br>Undefined; in practice the ability to respond sensibly in TL to TL questions | • COMPREHENSION<br>The ability to report accurately in English what is understood from a TL conversation with a native speaker; excludes non-interactive comprehension, such as understanding a media presentation. |
| ----- | • STRUCTURAL PRECISION<br>Morphology<br>Some aspects of syntax (grammaticality)<br>Phonology |
| • GRAMMAR<br>Morphology<br>Syntax | ----- |
| • ACCENT | ----- |
| ----- | • DISCOURSE COMPETENCE<br>Some aspects of syntax<br>Some aspects of semantics (text construction; organization and clarity of presentation)<br>Conversational management skills |
| • VOCABULARY<br>Breadth of lexicon | • LEXICALIZATION<br>Breadth of lexicon<br>Some aspects of semantics (appropriateness of encoding, both linguistic and socio-cultural) |
| • FLUENCY<br>Rate of speech | • FLUENCY<br>Continuity of speech<br>Accessibility of language |



BEST COPY AVAILABLE

BEST COPY AVAILABLE

2. The traditional performance standards--the degrees of competence associated with each factor--constituted a scale from "lack/error-ridden" to "mastery/error-free". Strong emphasis was given to correctness of form. The revised performance standards shift the emphasis to function, ie, to an assessment of effectiveness. The revised scale of performance standards defines three degrees of "difficulty" and three degrees of "success":

| | |
|---|---|
| BLOCKING | Examinees whose performance on a given factor is such that conversation frequently comes to a halt are rated "blocking" on that factor. |
| DYSFUNCTIONAL | Performance is such that interlocutor must strain to understand. |
| INTRUSIVE | Performance is comprehensible but causes distraction or annoyance to the interlocutor. |
| ACCEPTABLE | Performance is satisfactory but noticeably imperfect. |
| SUCCESSFUL | Performance is more than satisfactory, enabling examinee to succeed fully at even the most complex test tasks. |
| SUPERIOR | Performance is similar to that of a well-informed and articulate bilingual (Note: NOT necessarily a native speaker) |

3. The traditional test format--the testing events themselves--involved a native speaker interviewing the examinee. Conversations were to a large extent tester-led, and obliged the tester to provoke the examinee into demonstrating control (or lack of control) of particular features of the language which were considered markers of certain levels. The revised format standardizes the testing events for all tests in all languages by stipulating three task types:

- Task-type CNE -- Social/personal conversation

  A chat between tester and examinee on biographical topics initiated by the tester.

- Task-type TWO -- Interviewing

  Given a choice of topics (presented in written English), the examinee interviews the tester in the TL, and reports point-by-point (in oral English) on the tester's views.

- Task-type THREE -- Briefing

  Examinee briefs the tester in TL on information contained in one of three (written English) background papers, and responds to tester's TL questions and comments.

BEST COPY AVAILABLE